A comparison of position-specific score matrices based on sequence and structure alignments

ANNA R. PANCHENKO AND STEPHEN H. BRYANT

Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

(RECEIVED May 29, 2001; FINAL REVISION November 6, 2001; ACCEPTED November 6, 2001)

Abstract

Sequence comparison methods based on position-specific score matrices (PSSMs) have proven a useful tool for recognition of the divergent members of a protein family and for annotation of functional sites. Here we investigate one of the factors that affects overall performance of PSSMs in a PSI-BLAST search, the algorithm used to construct the seed alignment upon which the PSSM is based. We compare PSSMs based on alignments constructed by global sequence similarity (ClustalW and ClustalW-pairwise), local sequence similarity (BLAST), and local structure similarity (VAST). To assess performance with respect to identification of conserved functional or structural sites, we examine the accuracy of the three-dimensional molecular models predicted by PSSM-sequence alignments. Using the known structures of those sequences as the standard of truth, we find that model accuracy varies with the algorithm used for seed alignment construction in the pattern local-structure (VAST) > local-sequence (BLAST) > global-sequence (ClustalW). Using structural similarity of query and database proteins as the standard of truth, we find that PSSM recognition sensitivity depends primarily on the diversity of the sequences included in the alignment, with an optimum around 30–50% average pairwise identity. We discuss these observations, and suggest a strategy for constructing seed alignments that optimize PSSM-sequence alignment accuracy and recognition sensitivity.

Keywords: Profile search; protein structure alignment; alignment accuracy

Due to the success of genome sequencing efforts many proteins are now characterized only by sequence, with no experimental identification of their three-dimensional structure or function. Methods for pairwise sequence alignments may be used to infer structure and function by homology, but they may fail to detect distant evolutionary relationships in the "twilight zone" of sequence similarity. To improve sensitivity, methods based on residue conservation patterns within protein families have been developed (Gribskov et al. 1987; Eddy 1996; Hughey and Krogh 1996; Altschul et al. 1997; Karplus et al. 1997; Neuwald et al. 1997). The profile-search method PSI-BLAST (Altschul et al. 1997), for example, performs a database search by detecting increasingly divergent members of a given family in consecutive iterations. Detection of new family members is based on a position-dependent scoring matrix (PSSM) (Gribskov et al. 1987) derived initially from sequences aligned to a single-sequence query, or a "seed" alignment of previously known family members.

The sensitivity of a PSSM with respect to identification of divergent family members depends on the seed alignment used to construct it. Perfect discrimination between homologous and nonhomologous sequences can be achieved only when the PSSM is at once informative enough for specific recognition, and at the same time based on sequences that encompass the overall diversity of a protein family. As more and more diverse sequences are included in a seed alignment, the accuracy of that alignment may furthermore become an issue. Use of stringent gap penalties may cause misalignment of residues forming a conserved functional

Reprint requests to: Stephen H. Bryant, Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, Room 8N805, 8600 Rockville Pike, Bethesda, MD 20894; e-mail:bryant@ncbi.nlm.nih.gov; fax: (301) 435-7794.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.19902.

site, for example, if those residues are flanked by insertions or deletions. Inaccurate seed alignments may in this way introduce noise and unnecessarily dilute the information content of the PSSM.

Protein three-dimensional structure is remarkably stable with respect to sequence divergence, such that even the most distant relatives within a protein family exhibit the same overall topology and architecture. This resiliency allows proteins to gradually evolve into families with some variety of functional sites and functions, based on the same overall scaffold (Rost 1997; Holm 1998; Murzin 1998). Although it has been noted that conservation of structural features also decreases with evolutionary distance (Chothia and Lesk 1986; Hubbard and Blundell 1987; Flores et al. 1993; Russell and Barton 1994; Wood and Pearson 1999), it nonetheless seems possible that structural alignments may prove a useful source of seed alignments for PSSM construction. Alignments based on a conserved structural scaffold may accurately identify conserved sequence features and/or functional sites, even when overall sequence similarity is low.

The use of structure alignments for PSSM construction has been investigated previously by Sternberg and colleagues (Kelley et al. 1999, 2000). These investigators initiated database searches by merging the PSSMs of different sequence-similar subfamilies, based on structure–structure alignments for representatives of these subfamilies. This procedure aims to detect relationships between protein families that are not obvious from the component PSSMs individually. These investigators showed that combined PSSMs yielded recognition rates similar to the starting PSSMs, which can be explained by the extremely low similarity between the structurally similar subfamilies they considered (Kelley et al. 1999). Further analysis showed that some PSSMs constructed by this approach could indeed detect members of divergent subfamilies (Kelley et al. 2000). The investigators attributed this to the "mosaic" nature of the combined PSSMs, which simultaneously encode the sequence motifs characteristic of two or more sequence-dissimilar subfamilies.

Here we focus on a different application of structural alignments in PSSM construction. We compare the performance of PSSMs derived from seed alignments based on different sequence–sequence alignment algorithms to those based on structure–structure alignment. The seed alignments are in each case based on exactly the same protein sequences, and it is only their alignment per se, not the diversity of family members included in the alignment, that we vary in the experiments. Our intention is to measure whether and to what extent PSSM performance improves when seed alignments are based on the shared three-dimensional scaffold detected by structure–structure superposition, as opposed to the residue–conservation patterns detected by sequence–sequence comparison.

To assess PSSM performance we employ a test set where the 3-dimensional structures of both database and PSSM-template proteins are known. We may thus measure the accuracy of a PSSM-sequence alignment as that of the three-dimensional molecular model implied by that alignment and the known structure of the PSSM-template protein. We use one of the numerical measures of molecular model accuracy developed for the CASP structure–prediction competitions (Moult et al. 1997), contact specificity (Marchler-Bauer and Bryant 1997). We also measure recognition sensitivity of PSSMs based on seed alignments calculated by these different methods. To do so we examine the fraction of structurally similar proteins in the known-structure database that are identified with a significant PSI-BLAST E-values (Altschul et al. 1997).

In agreement with earlier results (Kelley et al. 1999, 2000), we find that use of structural alignments in PSSM construction has a modest effect on search sensitivity. We find a much greater effect on the accuracy of the PSSM-sequence alignments. When structural alignments are used to build the seed alignment, molecular models derived from PSSM-sequence alignments are in significantly better agreement with the known structure of the modeled proteins. We thus suggest that PSSMs derived from structural alignments may be most useful for accurate detection of the core-structure scaffold characteristic of a protein family and for annotation of functional sites associated with it.

Results

Measurement of PSSM-sequence alignment accuracy

The accuracy of a PSSM-sequence alignment may be understood as the fraction of residues from the PSSM-template protein that are correctly mapped to homologous residues in the aligned sequence. Because a similar spatial arrangement is a necessary condition for identification of homologous residues, alignment accuracy may be measured as the fraction of interresidue contacts (or sites) in the molecular model predicted by a PSSM-sequence alignment that are indeed present in the known three-dimensional structure of that protein. This is the quantity we measure as contact specificity (Marchler-Bauer and Bryant 1997). We emphasize that this metric is based on correct prediction of interresidue distances, not comparison to a true or reference alignment. We note in particular that the VAST structure–structure alignment algorithm (Gibrat et al. 1996), which we evaluate as a method of seed alignment construction, is not used as a standard of truth for evaluation of PSSM-sequence alignment accuracy.

To illustrate the contact specificity metric, we plot in Figure 1 average values obtained when molecular models based on VAST structure–structure alignments of PSSM-template and database proteins take the place of PSSM-

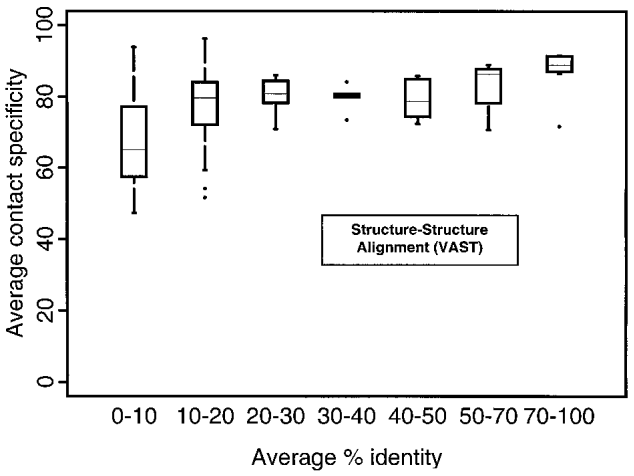

Fig. 1. Average contact specificity for molecular models predicted by structure–structure alignments of test-set sequences with structurally similar neighbors. The neighbors included in the averages are (as in Fig. 2) those detected by PSSMs from all four seed alignment methods we consider, with test-set domains grouped according to ranges of seed-alignment diversity. Results are shown as a boxplot (Chambers 1998), displaying the range of contact specificity values observed for each seed-alignment diversity range. The central line in each box shows the median contact specificity, the upper and lower boundaries of the box show the upper and lower quartiles, and the vertical lines extend to a value 1.5 times the interquartile range. Outlier values beyond these ranges are shown as individual points.

sequence alignments. These alignments may be understood as the most accurate one might expect from any PSSM-sequence alignment method, in that knowledge of the three-dimensional structures of the template and database proteins, rather than the template PSSM and database sequence, has been used in calculation of the alignment. One may see that median contact specificity values range from around 80%, for models based on VAST alignments of protein pairs with high sequence (and structure) similarity, to around 70%, for models based on structural alignment of protein pairs with lower sequence (and structure) similarity. Contact specificity does not reach 100%, because the structures of the query and template proteins are never identical to one another. Contact specificity values around 70% correspond to molecular models with root mean square superposition residuals (for polypeptide backbone atoms, when compared to the true structure) of around 3 Angstroms (not shown).

Alignment accuracy varies with seed alignment type

In Figure 2 we plot average contact specificity for molecular models based on PSSM-sequence alignments, for PSSMs from seed alignments calculated by different methods. We consider PSSMs from seed alignments calculated by local-structure (VAST), local-sequence (BLAST), and global-sequence (ClustalW and ClustalW-pairwise) comparison methods. We note that in each case (and in Fig. 1) values are averaged only over those database sequences detected by the PSSMs from all four seed alignment algorithms. Differences in average contact specificity thus reflect differences in PSSM-sequence alignment accuracy as a function of the seed alignment method, not differences in the sequence neighbors that have been included in averaging. We note that removing neighbors that are very similar to the template sequence (more than 50% identity) does not have a significant effect on the results shown.

As can be seen from Figure 2a–d, there are obvious differences in the PSSM-sequence alignment accuracy for PSSMs derived from different seed alignments. Above 50% average pairwise identity in the seed alignment, median contact specificity is near 80% for all alignment algorithms. Judging by comparison to structure–structure alignment results shown in Figure 1, these alignments are about as accurate as possible. Below 50% average pairwise identity in the seed alignment, however, alignment accuracy varies with the pattern local-structure (VAST) > local-sequence (BLAST) > global-sequence (ClustalW-pairwise) > global-sequence (ClustalW). PSSM-sequence alignments for PSSMs based on seed alignments from VAST are nearly as accurate as the corresponding structure–structure alignments, near 70% contact specificity, even when average pairwise identity in the seed alignment is 20% or lower. It appears that PSSMs from structure-based seed alignments better represent those sequence features that correspond to the conserved structural scaffold of a protein family.

To directly compare PSSM-sequence alignment accuracy for PSSMs from different seed alignment methods, we plot in Figure 3 average contact specificity for molecular models based on PSSMs from seeds by sequence comparison (BLAST, ClustalW, and ClustalW-pairwise) versus PSSMs from seed alignments by structure comparison (VAST). We note that the averages include all hits, not just those detected by PSSMs from all four seed alignment methods, and that average contact specificity is somewhat lower than in Figure 2 for this reason. It is apparent from Figure 3 that nearly all points fall below the diagonal, indicating that average PSSM-sequence alignment accuracy for nearly all sequences is greater when seed alignments are based on structure comparison. The only exceptions to the pattern are a few proteins where BLAST seed alignments lead to somewhat greater contact specificity; these are cases where the BLAST seed alignment is shorter than the VAST seed alignment, focusing on a highly similar region. It is striking that contact specificity sometimes rises from near zero, with PSSMs based on sequence alignments, to values over 70%, with PSSMs based on structure alignments.

Contact specificity measures correctly predicted interresidue contacts relative to the total number of predicted contacts. Thus, one might expect that PSSMs from seed alignments based on local sequence comparison, which focus on

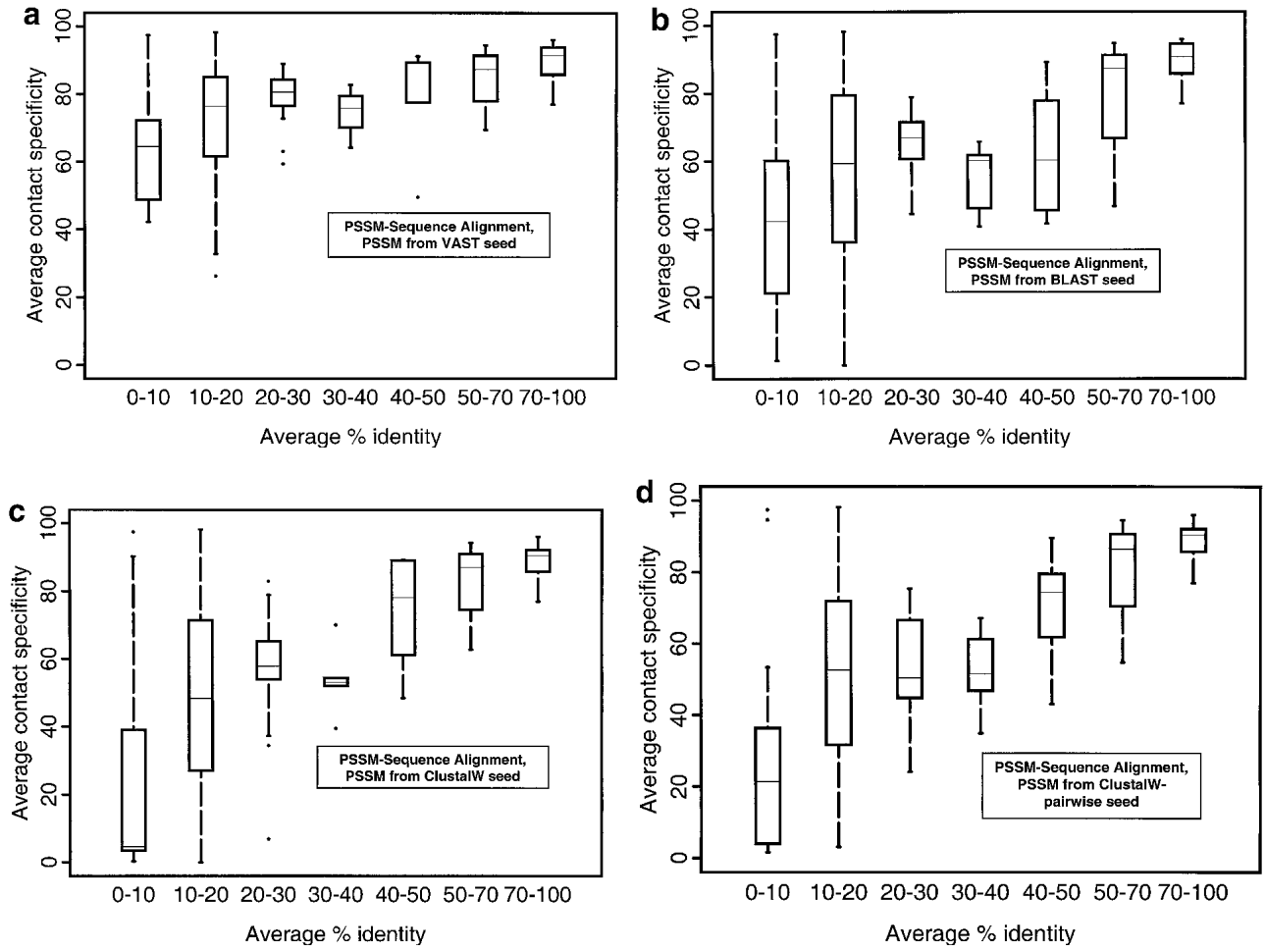

Fig. 2. Average contact specificity for molecular models predicted by PSSM-sequence alignments of test-set domains with structurally similar neighbors. PSSMs are calculated from seed alignments by VAST (*a*), BLAST (*b*), ClustalW (*c*), and ClustalW-pairwise (*d*). Test-set domains are grouped into ranges of seed-alignment diversity, based on average pairwise identity among all sequences in the seed, calculated via the VAST alignment of each sequence to the test-set domain. For purposes of comparison between different methods contact specificity is averaged only over those neighbor sequences identified with PSI-BLAST E-value < 0.01 by all four types of PSSM.

short, sequence-similar segments, may tend to score better under this metric than do PSSMs from global sequence comparison. Data in Table 1 show that PSSM-sequence alignments for PSSMs from BLAST seeds indeed tend to be shorter than alignments for PSSMs from ClustalW seed alignments. It thus seems likely that length differences to some extent account for the differences in accuracy seen in Figure 2 for local versus global sequence–sequence alignment. Data in Table 1 show that PSSM-sequence alignments for PSSMs from VAST seeds are nearly as long as those for PSSMs from ClustalW seed alignments, however. Differences in alignment length thus cannot account for the differences in PSSM-sequence alignment accuracy seen for sequence- versus structure-based seed alignments. For seed alignments in the 20–30% identity bin, for example, average PSSM-sequence alignment length is 90 residues for PSSMs from VAST seeds and 92 residues for PSSMs from ClustalW seed alignments (a difference of 2%), while median contact specificity values in Figure 2 are 80 and 50% respectively (a difference of 37%). We have also examined contact sensitivity, the fraction of correctly predicted contacts relative to all contacts in the structure of the database protein (Marchler-Bauer and Bryant 1997). Consistent with the above interpretation, we find that average contact sensitivity is greatest for PSSM-sequence alignments for PSSMs from VAST seed alignments (not shown).

In previous analyses of sequence alignment accuracy, Thompson et al. (1999) and Sauder et al. (2000) noted that in comparisons where there are many insertions and deletions, global alignment algorithms may be forced to align segments of nonhomologous residues that do not share structural similarity. Here, we find that PSSM-sequence

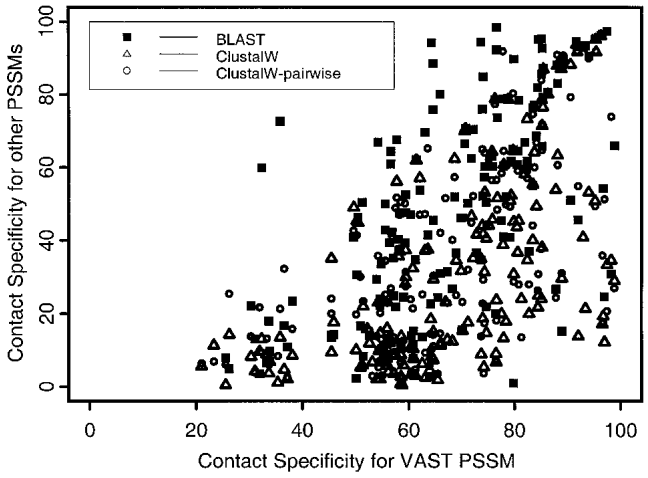

Fig. 3. Average contact specificity for molecular models predicted by PSSM-sequence alignments of test-set domains with structurally similar neighbors. Contact specificity is averaged separately over all models predicted by sequence-PSSM alignments from BLAST, ClustalW, and ClustalW-pairwise seed alignments. These values are plotted against the average contact specificity for models of the same test-set domain predicted by sequence-PSSM alignments from VAST seed alignments.

alignment accuracy, for PSSMs from seed alignments involving such low-similarity sequences, is lower for PSSMs from seeds based on global sequence alignment (ClustalW or ClustalW-pariwse) compared to PSSMs from seed alignments based on local sequence alignment (BLAST). Although Thompson et al. and Sauder et al. used different test sets, these observations seem consistent, and suggest that the PSSM-sequence alignments we evaluate reproduce the accuracy of the underlying seed alignments. The improved alignment accuracy we find for PSSMs from seeds based on structure–structure alignment (VAST) similarly suggests that PSSM-sequence alignments can also reproduce the accuracy of structure-based seed alignments.

Measurement of PSSM recognition sensitivity

The test set we employ in these experiments is based on 172 proteins of known structure, each of which is structurally similar to a large and diverse group of other known structures (see Materials and Methods). Each of these test-set proteins is used as a template sequence for calculation of PSSMs from seed alignments calculated by the different methods we evaluate, and our basic measure of PSSM sensitivity is simply the fraction of the structure neighbors of the test-set protein recognized with a significant PSI-BLAST E-value. To examine the effect on PSSM recognition sensitivity of increasing diversity among sequences in the seed alignment, we furthermore assign each test set protein to a particular range of seed alignment diversity. Because this assignment is based in part on the availability of structure neighbors in that diversity range, the proportion of structure neighbors with sequence similarity sufficient for recognition by PSSM-sequence comparison may vary among seed-alignment diversity ranges.

To correct for any differences in diversity in the structure–neighbor set we employ as a standard of truth, we also examine the number of structure neighbors recognized relative to the number we would expect to recognize, if PSSM performance were equivalent across different seed-alignment diversity ranges. We estimate the number of structure neighbors we would expect to recognize simply as those with greater than 12% identical residues in VAST structure–structure alignment. This threshold was previously identified as the point where homologous structure neighbors, related by descent from a common ancestral gene, begin to outnumber "analogous" structure neighbors, which may reflect convergent evolution (Matsuo and Bryant 1999). We emphasize that comparison of PSSM performance for different seed alignment methods, within a given seed alignment diversity range, is not affected by this correction, because the total number of structure neighbors recognized is simply divided by a constant. The correction is useful for comparison of PSSM sensitivity across seed-alignment diversity ranges.

Table 1. *Average PSSM-sequence alignment length for PSSMs from seed alignments by VAST, ClustalW, Blast, and ClustalW-pairwise, for ranges of seed-alignment diversity*

	0–10	10–20	20–30	30–40	40–50	50–70	70–100
Vast	105	84	90	97	133	152	107
Clustal	122	89	92	102	133	152	106
Blast	78	67	76	84	125	144	105
Pair-clustal	124	88	91	102	132	150	106

Seed-alignment diversity is expressed as the average percentage of identical residues among all pairwise comparisons of sequences in the seed alignment, calculated via the VAST alignment of each with the sequence of the test-set domain.

Diversity of seed alignments determines recognition sensitivity

In Figure 4 we plot structure neighbor recognition rates, for PSSMs calculated from seed alignments by different methods, for ranges of seed-alignment diversity. One pattern apparent from Figure 4 is that even the most sensitive PSSMs still miss many similarities detected by structure–structure comparison, such that average overall sensitivity is only about 20%. This reflects the difficulty of the test set, where many structure neighbors have no detectable sequence similarity. Using a different set of structure neighbors as the standard of truth, Brenner and colleagues similarly found that only a small fraction of structure neighbors may be detected by sequence–sequence or PSSM-sequence comparison methods (Brenner et al. 1998).

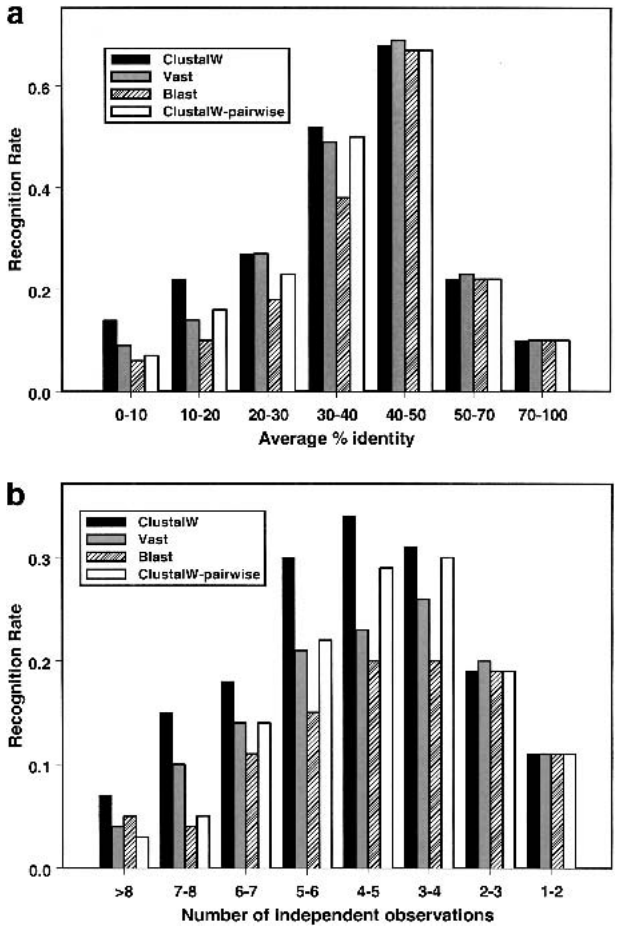

Fig. 4. PSSM recognition sensitivity for ranges of average percent identity (*a*) and average number of independent observations (*b*) of sequences in seed alignments. Each bar represents the mean recognition rate for PSSMs based on seed alignments by the methods indicated, for the indicated range of seed-alignment diversity. Domains in the test set are listed by their PDB code (lower case), chain identifier (if applicable, upper case) and domain identifiers (numeric, starting with 1 for each chain): 1a2zA, 1a66A, 1a6cA1, 1a8z, 1a96C, 1aac, 1aazA, 1abe 2, 1abrB2, 1adoA, 1afj, 1ah1, 1aizA, 1aj0, 1ajsA2, 1ak5, 1aozA1, 1aq0A1, 1ash, 1atzB, 1auyB, 1auz, 1av6A2, 1avc 1, 1aw5 1, 1aym3, 1be1, 1bebA, 1bf5A2, 1ble, 1bmdA1, 1bmtA2, 1bmv1, 1bmv21, 1bmv22, 1bovA, 1boy 1, 1boy 2, 1bp3B1, 1bp3B2, 1bquB1, 1bquB2, 1bslB, 1c25, 1cdh, 1cen, 1cfb 1, 1cfb 2, 1cpcB1, 1ctn 1, 1cto, 1cwpB, 1dcpC, 1dhr, 1din, 1dpgA1, 1dpmA, 1e2b, 1eayD, 1ebpA1, 1ebpA2, 1eca, 1eceA, 1edg, 1edhA2, 1eft 1, 1efvA1, 1efvB1, 1epaB, 1epnE2, 1f13A1, 1f13A4, 1fem, 1fivA, 1fmtA1, 1fnf 2, 1fnf 3,1fod1, 1fts 2, 1grx, 1hbg, 1hcd, 1hjrA, 1hnf 1, 1hnf 2, 1hoe, 1hstA, 1IdaA, 1itbB1, 1itbB3, 1ithA, 1jdbK5, 1jdbK8, 1jer, 1jli, 1jlxA1, 1jlxA2, 1jrhI, 1kb5B, 1ksr, 1kte, 1lea, 1lki, 1nal11, 1neu, 1nfkA2, 1occB1, 1ofgA1, 1opc 2, 1ordA1, 1pamA3, 1pdo, 1pii 1, 1pii 2, 1pnt, 1pysB7, 1qapA2, 1rcb, 1rhoA, 1ris, 1rvv1, 1scuA1, 1scuB3, 1sfe 2, 1sftA2, 1soxA3, 1sro, 1stmA, 1svb 3, 1tbgA2, 1tde 2, 1tdj 2, 1ten, 1uag 1, 1uag 3, 1udiI, 1vcaA1, 1vcaA2, 1wab 1, 1who, 1xan 3, 1xbrA, 1yub 1, 1yveI1, 1zxq 1, 1zxq 2, 2awo, 2dldA2, 2dri 2, 2fmr, 2gdm, 2gmfA, 2I1b, 2ila, 2mnr 2, 2ncm, 2pgd 1, 2pii 1, 2rspA, 2sas 2, 2stv 1, 2tmdA3, 2trxA, 2u1a, 2wbc, 3btoA2, 3chy, 3inkC, 3ullA, 5p21, 5ptp, 1tde 1.

In Figure 5 we plot relative recognition rates, where we divide the number of structure neighbors recognized by the number one might expect to recognize for that seed-alignment diversity range. It is apparent from both Figures 4 and 5 that PSSM recognition sensitivity varies strongly with seed-alignment diversity, as judged by either relative or absolute recognition rates. As one may see from Figures 4a or 5a, optimal PSSM recognition sensitivity is obtained from seed alignments with intermediate diversity, in the range of 30–50% average pairwise identity. The optimal range of seed alignment diversity may alternatively be expressed as between three and five different residue types per aligned site (number of independent observations), as shown in Figures 4b and 5b, or as 0.5–0.6 bits of information per site in the PSSM (not shown). These maxima stay in the same range if a more conservative threshold of $E \leq 10e\text{-}5$ is used in the PSI-BLAST search (not shown). Several investigators have suggested that recognition sensitivity should in general vary with the degree of divergence

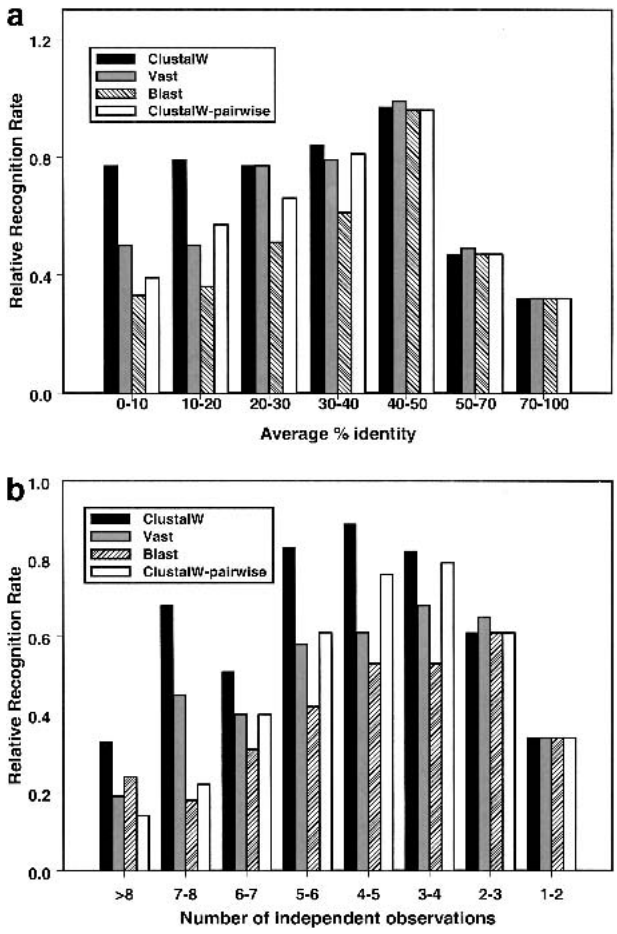

Fig. 5. Relative recognition rates for PSSMs derived from different seed alignments for ranges of average percent identity (*a*) and average number of independent observations (*b*). Each bar represents a ratio of the number of structure neighbors recognized and the number of structure neighbors one might expect to recognize for each seed-alignment diversity range (see Results section).

of the aligned sequences used to calculate the PSSM (Park et al. 1998; Aravind and Koonin 1999; Salamov et al. 1999; Rychlewski et al. 2000), and the current analysis supports this conclusion.

We note that recognition specificity (the fraction of proteins identified in the PSI-BLAST search that are indeed structurally similar) is comparably high for the different diversity ranges and seed alignment methods shown in Figures 4 and 5. The improved recognition sensitivity for seed alignments with 30–50% pairwise identity does not come at the expense of decreased specificity, in other words. Actual specificity values for all experiments range from 96–99%. Specificity values slightly below 99% (the value expected from the PSI-BLAST E-value cutoff we employ) seem to arise from a few protein pairs missing in the structure neighbor set we use as the standard of truth. Some flexible and/or disordered proteins are not detected as structure neighbors, and are absent from our standard of truth, even when they have significant sequence similarity (not shown).

It is also apparent from Figure 5 that for seed alignments below 30% average pairwise identity or above an average of four residue types per aligned position there is some variation of recognition sensitivity with alignment type. Recognition sensitivity decreases with the pattern global-sequence (ClustalW) > local-structure (VAST) ~ global-sequence (ClustalW-pairwise) > local-sequence (BLAST), although it is also apparent that this variation is smaller than the effect of seed-alignment diversity. Global sequence alignments (ClustalW and ClustalW-pairwise) presumably perform better than local sequence alignment (BLAST) because the domain pairs in the test set are indeed globally similar, and the PSSMs formed from these longer alignments more sensitive, an effect observed before (Thompson et al. 1999; Notredame et al. 2000). As can be seen from comparing the same alignment algorithm (ClustalW) used in two different ways, the multiple alignment method ClustalW performs better than ClustalW used in a pairwise fashion. Indeed, in situations where the seed alignment is constructed from different subfamilies, sequence motifs common to a subset of neighbors but not present in the test-set domain can only be aligned correctly using multiple alignment tools.

To examine the effect of gap content on PSSM recognition sensitivity we compare the recognition sensitivity of PSSMs derived from alignments with approximately the same fraction of gaps, as shown in Table 2. Only alignments with average sequence identity below 30% are included. As can be seen from Table 2, seed alignments with more gaps produce less sensitive PSSMs for all alignment algorithms, presumably because the sequences in these alignments are among the most diverse. Interestingly, for seed alignments containing equal fractions of gaps, the local alignment methods perform as well as the global alignment methods, and the local-structure (VAST) method gives the most sensitive PSSMs. The local alignments are presumably more accurate in the regions they have aligned, and when the alignment is as extensive as that from ClustalW, containing an equal fraction of gaps, PSSM sensitivity is comparable.

Table 2. *Average recognition sensitivity for PSSMs based on seed alignments calculated by different methods, for seed alignments with different gap content*

Gap fraction	VAST	ClustalW	pairClustalW	Blast
≥0.25	0.12	0.10	0.06	0.10
<0.25	0.27	0.23	0.18	0.22

Gap content is expressed as the mean number of gap characters per column in the seed alignment. Only seed alignments with less than 30% average pairwise identity among sequences in the seed alignment are included in this analysis.

Discussion

In the present experiments we consider a difficult test for PSSM-based search methods, recognition of structure neighbors where sequence conservation may be very low. We find, in agreement with earlier work (Kelley et al. 2000), that it is not generally possible to produce a single PSSM capable of detecting all these neighbors, no matter what method of seed alignment construction we consider. Including overly diverse sequences in the seed alignment, even if aligned by structure comparison, simply tends to dilute the information content of the PSSM. We find that there is an optimal range of sequence diversity to consider in making seed alignments. Seed alignments with 30–50% of average percent identity or three to five average amino acid types per aligned position detect a greater fraction of structure neighbors than do seeds with either greater or lesser diversity. This suggests a strategy for constructing the minimal set of PSSMs needed to recognize members of a diverse structural family: One should divide the family into subfamilies containing sequences with 30–50% pairwise identity, and construct PSSMs for each.

Although the method of seed alignment construction seems to have little effect on search sensitivity, we find just the opposite result with respect to the accuracy of PSSM-sequence alignments. When we examine the accuracy of the 3D molecular model implied by the PSSM-sequence alignment, we find that structure-based seed alignments produce PSSMs that better detect and reproduce the conserved core structure characteristic of a protein family. This effect is most pronounced for PSSM-sequence alignments where the PSSM is derived from seed alignments of very diverse sequences, but it is apparent from the data presented in Figures 2 and 3 that use of structure-based seed alignments rarely leads to a decrease in alignment accuracy. Because the PSI-BLAST algorithm in general tends not to start or extend HSPs (high-scoring segment pairs) where there were

gaps in the seed alignment, it is perhaps not surprising that PSSM-sequence alignments, for PSSMs from structure-alignment seeds, which have no gaps within core elements, better reproduce the conserved structural scaffold. This observation suggests a simple strategy for improving the accuracy of PSSM-sequence alignments and the reliability of annotations derived from them: whenever possible, use structure alignments as seeds for PSSM construction.

Last, it is interesting to ask why PSSMs from seed alignments by structure–structure comparison can have a strong effect on PSSM-sequence alignment accuracy, but relatively little effect on recognition sensitivity. This result at first appears paradoxical, because one might suppose that more accurate seed alignments, if this indeed accounts for the effects we observe, might lead to improvements in both sensitivity and accuracy. The explanation may simply be that characteristic sequence motifs, sufficient for sensitive recognition of family members, are in general well detected by all of the seed alignment methods we consider. The improved alignment accuracy we observe for PSSMs from structure–structure alignment suggests that they may better represent additional regions of similarity, where sequence similarity is too weak for accurate alignment by sequence comparison algorithms. Precisely because sequence similarity in these regions is weak, however, one would not expect a large change in the information content of PSSMs calculated from these seed alignments, or a large change in recognition sensitivity. We can thus suggest that the primary effect of using structure–structure alignments in PSSM construction will be to improve PSSM-sequence alignment accuracy, and as a consequence, the accuracy of annotation transfer from known family members to new sequences detected by PSSM-based search tools. If one knows the locations of active site residues in a structure-based seed alignment, for example, one may expect that a PSSM-sequence alignment based on that seed may more accurately identify the homologous active site residues in new sequences.

Materials and methods

Representative domains

To obtain a representative set of template structures for use in seed alignments we selected 2900 sequence-dissimilar domains from the Protein Data Bank (PDB) (Berman et al. 2000). The set was constructed by single-linkage clustering based on a BLAST E-value of 10e-7 or less, as described previously (Matsuo and Bryant 1999). Domain boundaries were taken from MMDB (Marchler-Bauer et al. 1999), the structure database that is distributed with Entrez (http://www.ncbi.nlm.nih.gov/Entrez/). Domains in MMDB are identified using a compactness algorithm similar to that of Holm and Sander (1994), as described previously (Madej et al. 1995). A similar set was employed in a recent analysis of threading sensitivity and accuracy (Panchenko et al. 2000). A listing is available at http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html.

We omit from the test set sequence-discontinuous domains (except for 1TDE 1), domains that have less than five structure neighbors within the test set, and domains longer than 250 residues. The requirement for five or more sequence-dissimilar structure neighbors is restrictive, because there are few protein families for which this many structures are known, and it reduces the test set to a total of 172 domains. Structure neighbors are again taken from the database distributed with Entrez (http://www.ncbi.nlm.nih.gov/Entrez/), as identified by the VAST structure comparison algorithm (Madej et al. 1995; Gibrat et al. 1996). We note that structural similarities detected by the VAST algorithm have previously been compared to the SCOP classification (Murzin et al. 1995; Matsuo and Bryant 1999; Przytycka et al. 1999). According to SCOP (Murzin et al. 1995), this set of 172 domains includes five different classes, 50 folds, and 73 superfamilies, and by this criterion may be considered a diverse sample of protein domains.

Homologous structure neighbors

We include in seed alignments only sequences from homologous structure neighbors. Homologous structure neighbors for each test-set domain are chosen from the complete structure neighbor set on the basis of significant sequence similarity or extensive structural similarity consistent with descent from a common ancestral gene. We examine all VAST neighbors within MMDB, but include only those neighbors classified as belonging to the same homologous superfamily by the authors of SCOP (Murzin et al. 1995), or, when no SCOP classification is available, with more than 12% sequence identity in the VAST structure alignment. These structure neighbors are then used to calculate the "homologous core substructure" (HCS) of the template domain, and any additional structure neighbors superimposing onto 90% or more of the HCS residues are recruited as additional homologous neighbors, as described previously (Matsuo and Bryant 1999).

Because we anticipate that the diversity of sequences in a seed alignment will have an effect on PSSM performance, we do not select all homologous structure neighbors of a given template domain when constructing seed alignments for PSSMs. Instead, we sample randomly among them, to select a subset of neighbors with a defined range of sequence similarity. We assign each template domain randomly to a range of sequence similarity, 0–10%, >10–20%, etc. We then randomly choose a homologous structure neighbors exhibiting this range of sequence similarity with respect to the template, iterating this process until at least 5 but no more than 50 neighbors are selected. If this process fails for a given template domain and target similarity range, due to insufficient structure neighbors, that domain is randomly exchanged for another from a different similarity range, and neighbor selection begun anew.

Construction of seed alignments

Structure alignments of the template domain with its homologous structure neighbors are taken directly from VAST alignments as described above. Pairwise sequence alignments between the template domain and these neighbors are calculated using the gapped BLAST algorithm (Altschul et al. 1997) and the ClustalW algorithm (Thompson et al. 1994), applied separately for the template and each neighbor sequence. To produce multiple sequence alignments we apply the ClustalW algorithm to the template and all selected neighbors. In this application ClustalW constructs the alignment progressively, grouping the most similar sequences into

aligned clusters and aligning larger and larger alignment clusters with one other (Thompson et al. 1994). As a result, we end up with four different types of seed alignment, one based on automatic structure–structure alignment (VAST) and three based on different automatic sequence–sequence alignment algorithms (BLAST, ClustalW-pairwise, and ClustalW), each run with default parameters suggested by the authors of the algorithm.

We derive PSSMs from seed alignments using the default method of PSI-BLAST. Aligned sequences are projected onto the template domain (insertions relative to the template are ignored) and a PSSM calculated using the pseudocount method described previously (Altschul et al. 1997). PSSMs derived from each type of alignment are used to initialize searches of a database consisting of nonidentical sequences from PDB. We perform only a single iteration of PSI-BLAST (Altschul et al. 1997), so as to avoid any modification of the seed-derived PSSM, and collect hits with E-values below 0.01. We emphasize that seed alignments based on VAST, BLAST, ClustalW-pariwise, and ClustalW methods contain exactly the same template and neighbor sequences, differing only in the algorithm that has been employed in building the alignment.

Fold recognition sensitivity and alignment accuracy

The VAST structure neighbors of each template domain provide a standard of truth for judging the sensitivity of PSSMs constructed from each type of seed alignment. Hits to structurally similar neighbors are considered true positives, misses of structurally similar neighbors are considered false negatives. We thus calculate fold recognition sensitivity as ($N^{\mathrm{tr.pos}}/N^{\mathrm{VAST.neigh}}$) and fold recognition specificity as ($N^{\mathrm{tr.pos}}/N^{\mathrm{PDB.hits}}$). Here, $N^{\mathrm{tr.pos}}$ is the number of nonidentical structure neighbors detected with E-value below a given threshold, $N^{\mathrm{VAST.neigh}}$ is the overall number of nonidentical structure neighbors for a given domain, and $N^{\mathrm{PDB.hits}}$ the total number of sequences (with known structure) that are identified with significant PSI-BLAST E-value. We note that these counts exclude any structure neighbors used in the seed alignments or any sequences identical to them. Counts may include sequences similar to those in the seed alignment, however, because our intention is to compare PSSM performance with respect to retrieval and accurate alignment of database sequences spanning a range of similarities with respect to the sequences in the seed alignment.

To evaluate the accuracy of the PSSM-sequence alignments we use the known structures of the template domains and of the database proteins identified by PSI-BLAST search. Using the known structure of the identified database protein as the standard of truth, we evaluate the accuracy of the molecular model implied by the PSSM-sequence alignment of the template domain with the sequence of that protein. We employ the numerical measure contact specificity, defined as the percent of nonlocal residue contacts in the predicted structure that are also present in the experimental structure (Marchler-Bauer and Bryant 1997): $\mathrm{ACSpc} = N^{\mathrm{cp}}/N^{\mathrm{p}}$. Here, N^{cp} is the number of nonlocal contacts (for residues separated along the chain by at least five peptide bonds and having C_α-atoms less than 8 Angstroms apart) that occur in both the molecular model implied by PSSM-sequence alignment and in the experimental structure of the database protein. N^{p} is the total number of nonlocal contacts in the predicted model. As in calculation of recognition sensitivity, family members present in the seed alignment are ignored in evaluation of alignment accuracy. In comparing accuracy of PSSM-sequence alignments for PSSMs derived from different seed alignments, we average contact specificity across the various database proteins identified by each PSSM.

Acknowledgments

We thank Aron Marchler-Bauer and Tom Madej for assistance with alignment accuracy evaluation, VAST neighbor calculations, and helpful discussions. We thank the reviewers for useful suggestions and the NIH Intramural Research Program for support.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Aravind, L. and Koonin, E.V. 1999. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.* **287:** 1023–1040.

Berman, H.M., Bhat, T.N., Bourne, P.E., Feng, Z., Gilliland, G., Weissig, H., and Westbrook, J. 2000. The Protein Data Bank and the challenge of structural genomics. *Nat. Struct. Biol. (Suppl.)* **7:** 957–959.

Brenner, S.E., Chothia, C., and Hubbard, T.J. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci.* **95:** 6073–6078.

Chambers, J.M. (1998). *Programming with data. A guide to the S language.* Springer-Verlag, New York.

Chothia, C. and Lesk, A.M. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5:** 823–826.

Eddy, S.R. 1996. Hidden Markov models. *Curr. Opin. Struct. Biol.* **6:** 361–365.

Flores, T.P., Orengo, C.A., Moss, D.S., and Thornton, J.M. 1993. Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* **2:** 1811–1826.

Gibrat, J.F., Madej, T., and Bryant, S.H. 1996. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6:** 377–385.

Gribskov, M., McLachlan, A.D., and Eisenberg, D. 1987. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci.* **84:** 4355–4358.

Holm, L. 1998. Unification of protein families. *Curr. Opin. Struct. Biol.* **8:** 372–379.

Holm, L. and Sander, C. 1994. Parser for protein folding units. *Proteins* **19:** 256–268.

Hubbard, T.J. and Blundell, T.L. 1987. Comparison of solvent-inaccessible cores of homologous proteins: Definitions useful for protein modelling. *Protein Eng.* **1:** 159–171.

Hughey, R. and Krogh, A. 1996. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *Comput. Appl. Biosci.* **12:** 95–107.

Karplus, K., Sjolander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L., and Sander, C. 1997. Predicting protein structure using hidden Markov models. *Proteins* **Suppl. 1:** 134–139.

Kelley, L.A., MacCallum, R.M., and Sternberg, M.J.E. 1999. Recognition of remote protein homologies using three-dimensional information to generate a position specific scoring matrix in the program 3D-PSSM. Proceedings of RECOMB, Lyon, France.

Kelley, L.A., MacCallum, R.M., and Sternberg, M.J. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299:** 499–520.

Madej, T., Gibrat, J.F., and Bryant, S.H. 1995. Threading a database of protein cores. *Proteins* **23:** 356–369.

Marchler-Bauer, A., Addess, K.J., Chappey, C., Geer, L., Madej, T., Matsuo, Y., Wang, Y., and Bryant, S.H. 1999. MMDB: Entrez's 3D structure database. *Nucleic Acids Res.* **27:** 240–243.

Marchler-Bauer, A. and Bryant, S.H. 1997. Measures of threading specificity and accuracy. *Proteins* **Suppl. 1:** 74–82.

Matsuo, Y. and Bryant, S.H. 1999. Identification of homologous core structures. *Proteins* **35:** 70–79.

Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K., and Pedersen, J.T. 1997. Critical assessment of methods of protein structure prediction (CASP): Round II. *Proteins* **Suppl. 1:** 2–6.

Murzin, A.G. 1998. How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8:** 380–387.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A

structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Neuwald, A.F., Liu, J.S., Lipman, D.J., and Lawrence, C.E. 1997. Extracting protein alignment models from the sequence database. *Nucleic Acids Res.* **25:** 1665–1677.

Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302:** 205–217.

Panchenko, A.R., Marchler-Bauer, A., and Bryant, S.H. 2000. Combination of threading potentials and sequence profiles improves fold recognition. *J. Mol. Biol.* **296:** 1319–1331.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284:** 1201–1210.

Przytycka, T., Aurora, R., and Rose, G.D. 1999. A protein taxonomy based on secondary structure. *Nat. Struct. Biol.* **6:** 672–682.

Rost, B. 1997. Protein structures sustain evolutionary drift. *Fold. Des.* **2:** S19–S24.

Russell, R.B. and Barton, G.J. 1994. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J. Mol. Biol.* **244:** 332–350.

Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A. 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **9:** 232–241.

Salamov, A.A., Suwa, M., Orengo, C.A., and Swindells, M.B. 1999. Genome analysis: Assigning protein coding regions to three-dimensional structures. *Protein Sci.* **8:** 771–777.

Sauder, J.M., Arthur, J.W., and Dunbrack, R.L., Jr. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins* **40:** 6–22.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Thompson, J.D., Plewniak, F., and Poch, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* **27:** 2682–2690.

Wood, T.C. and Pearson, W.R. 1999. Evolution of protein sequences and structures. *J. Mol. Biol.* **291:** 977–995.